Information Extraction Workflow for Digitised Entry-based Documents

Mohamed Khemakhem, Simon Gabay, Béatrice Joyeux-Prunel, Laurent Romary, Léa Saint-Raymond, Lucie Rondeau Du Noyer

Information Extraction Workflow for Digitised Entry-based Documents

Mohamed Khemakhem[1,2,3,4], Simon Gabay[5,6], Béatrice Joyeux-Prunel[6], Laurent Romary[2,3,7], Léa Saint-Raymond[8], and Lucie Rondeau Du Noyer[0]

[1]LittArts - UMR 5316, Grenoble
[2]Inria - ALMAnaCH - Automatic Language Modelling and ANAlysis Computational Humanities
[3]CMB - Centre Marc Bloch
[4]Université Paris Diderot, Paris
[5]Université de Neuchâtel
[6]Université de Genève
[7]BBAW - Berlin-Brandenburg Academy of Sciences and Humanities
[8]Ecole normale supérieure, Paris, DMA

January 2020

1 Program

It is a 4 hours workshop:

- 30 mn: introduction of the workflow and the two tools
- 30 mn: setting up the tools on the machines
- 2 h: Two Parallel groups. Goal: Training a model for a dictionary and a model for a catalogue
- 30 mn: towards a digitisation pipeline
- 30 mn: Q&A

2 Instructors

- Mohamed Khemakhem

- Simon Gabay
- Laurent Romary

3 Target Audience

Scholars and humanists with basic notions of command lines and familiar with the TEI standard.

4 Abstract

The massive retro-digitisation of legacy paper resources in the last decade, along with the constant growth of compiled unstructured digital text material, have created an unbalanced situation where the existent *ad hoc* techniques for exploiting such resources are unable to cover the important stream of emerging corpora. In this workshop we address this issue and present an exploratory workflow implemented in two state of the art infrastructures for Information Extraction (IE) from documents with entry-based structure and diverse content.

IE in Digital Humanities (DH) has always been a serious challenge for researchers dealing with modern or legacy text resources [1, 2]. *GROBID-Dictionaries*[1] is a project which has been launched to fill in this gap by accelerating the modelling and structuring of resources within the lexicography field. The first version of the machine learning infrastructure has been focused on structuring digitised dictionaries into TEI-compliant resources [3].

In *GROBID-Dictionaries*, the activation of cascading IE models follows an exploratory process based on the MATTER workflow [4]. Throughout a multi-stage annotation and curation process, a user of the tool discovers gradually the structure and the variation of the information in a target document.

[1] https://github.com/MedKhem/grobid-dictionaries

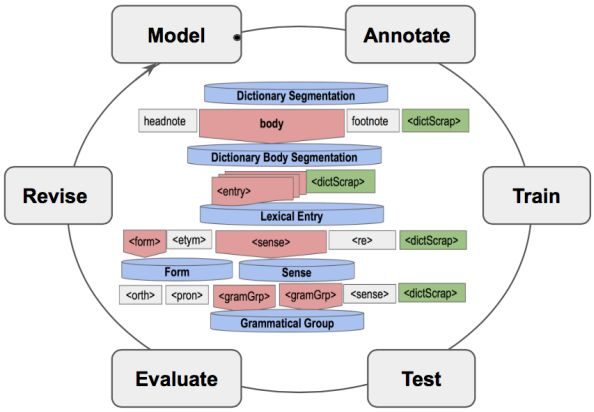
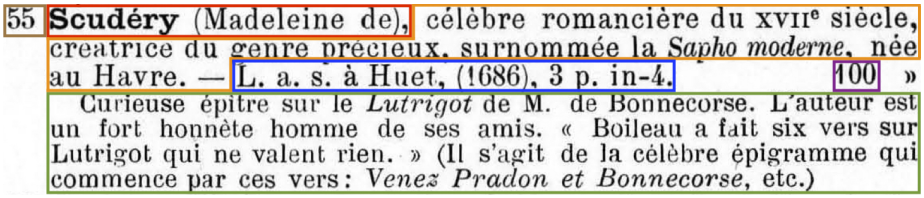


Figure 1: *GROBID-Dictionaries*'s MATTER Workflow

In addition, and despite being initially designed for structuring dictionaries, the implemented approach and the tool's pluggable models have shown enough flexibility to be applicable on a wider range of modern and legacy entry-based documents such as dictionaries [3, 5] (fig. 4), legacy address directories [6] (fig. 6), and large bibliographic collections [7].

The rising interest around the applicability of such an IE approach to certain entry-based documents [8, 9] and its fitness potential in larger document processing pipelines, triggered the creation of a second tool adapted for analysing and structuring a family of prints: *Catalogues* (fig. 5 & 7). *Grobid-Cat*[2] is a fork of *Grobid-Dictionaries* that remains faithful to the exploratory cascading approach with more optimised combination of parsing models (fig. 2) and more adapted TEI encoding (fig. 3).


55 **Scudéry** (Madeleine de), célèbre romancière du XVIIe siècle, creatrice du genre précieux, surnommée la *Sapho moderne*, née au Havre. — L. a. s. à Huet, (1686), 3 p. in-4. 100 »

Curieuse épître sur le *Lutrigot* de M. de Bonnecorse. L'auteur est un fort honnête homme de ses amis. « Boileau a fait six vers sur Lutrigot qui ne valent rien. » (Il s'agit de la célèbre épigramme qui commence par ces vers : *Venez Pradon et Bonnecorse*, etc.)

Figure 2: *RDA*, No 67 (March 1881), lot No 55.

[2]https://github.com/MedKhem/grobid-cat

```
<item n="55" xml:id="CAT_000037_e55">
   <num>55</num>
   <name type="author">Scudéry (Madeleine de),</name>
   <trait>
      <p>célèbre romancière du XVIIe siècle, créatrice du genre
      précieux, surnommée la Sapho moderne, née au Havre.</p>
   </trait>
   <desc>L.a.s. à Huet, (1686), 3 p. in-4. 100 </desc>
   <note>Curieuse épître sur le Lutrigot de M. de Bonnecorse.
   L'auteur est un fort honnête homme de ses amis.  Boileau
   a fait six vers sur Lutrigot qui ne valent rien.  (Il
   s'agit de la célèbre épigramme qui commence par ces vers:
   Venez Pradon et Bonnecorse, etc.)</note>
</item>
```

Figure 3: *RDA*, Output from Figure 1 processed with GROBID-Cat.

The proposed workshop represents an illustration of scholarly primitives practices employing a new generation of IE tools. Our goal is to familiarise interdisciplinary DH users with the MATTER workflow through two parallel hands-on sessions, each one dealing with a different category of digitised documents: dictionaries and catalogues. We rely on enhanced usability features [5] to ease the setup and the manipulation of the two systems for relaxed required IT skills. Besides presenting new powerful means for manipulating, exploring and extracting structured text from digitised material, the workshop gives DH researchers the opportunity to have an idea about the challenges of integrating such tools in digitisation pipelines. This includes providing background on the different factors interfering in the performance of machine learning models, such as information modelling, text encoding, OCR system choice and annotation consistency.

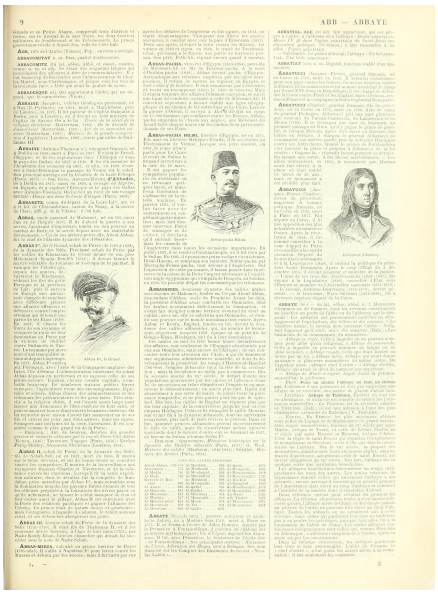


Figure 4: *Nouveau Larousse illustré: dictionnaire universel encyclopédique*, 1898, p 9

Figure 6: *Annuaire-almanach du commerce, de l'industrie, de la magistrature et de l'administration*, 1871, p. 200

Figure 5: *RDA*, No 37 (august 1873).

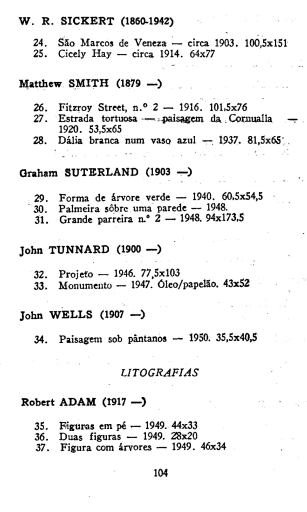


Figure 7: *I Bienal do museu de arte moderna de São Paulo*, 1951, p. 104

References

[1] Ian Gregory. Challenges and opportunities for digital history. *Frontiers in digital humanities*, 2014.

[2] José van Dijck. Big data, grand challenges: on digitization and humanities research. *KWALON*, 21:8 – 18.

[3] Mohamed Khemakhem, Luca Foppiano, and Laurent Romary. Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In *electronic lexicography, eLex 2017*, Leiden, Netherlands, September 2017.

[4] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications.* O'Reilly Media, 2012.

[5] Mohamed Khemakhem, Axel Herold, and Laurent Romary. Enhancing Usability for Automatically Structuring Digitised Dictionaries. In *GLOBALEX workshop at LREC 2018*, Miyazaki, Japan, May 2018.

[6] Mohamed Khemakhem, Carmen Brando, Laurent Romary, Frédérique Mélanie-Becquet, and Jean-Luc Pinol. Fueling Time Machine: Information Extraction from Retro-Digitised Address Directories. In *JADH2018 "Leveraging Open Data"*, Tokyo, Japan, September 2018.

[7] David Lindemann, Mohamed Khemakhem, and Laurent Romary. Retrodigitizing and Automatically Structuring a Large Bibliography Collection. In *European Association for Digital Humanities (EADH) Conference*, Galway, Ireland, December 2018.

[8] Lucie Rondeau Du Noyer, Simon Gabay, Mohamed Khemakhem, and Laurent Romary. Scaling up Automatic Structuring of Manuscript Sales Catalogues. In *TEI 2019: Book of Abstracts*, Graz, Austria, September 2019.

[9] Simon Gabay, Lucie Rondeau du Noyer, Mohamed Khemakhem, and Laurent Romary. Selling autograph manuscripts in 19th c. paris: digitising the revue des autographes. In *AIUCD 2020: Proceedings of the Ninth Annual Conference*, Milan, Italy, January 2020.